

# CLAIMS

I claim:

1. A cache coherency system for a shared memory parallel processing system including plurality of processing nodes, comprising: a multi-stage communication network for interconnecting said processing nodes; each said processing node including one or more caches for storing a plurality of cache lines; and a cache coherency directory which is distributed to each of said nodes for tracking which of said nodes have copies of each cache line.

2. A shared memory parallel processing system including a plurality of processing nodes, comprising:

a multi-stage communication network for interconnecting said processing nodes, said network including a plurality of self-routing switches cascaded into first, middle and last stages, each said switch including a plurality of switch inputs and a plurality of switch outputs, each of said switch outputs of each said switch coupled to a different switch input of others of said switches, switch outputs of said last stage switches including network output ports, and switch inputs of said first stage switches comprising network input ports;

660760-49546260

14           each processing node including:

15                 a network adapter for transmitting and receiving

16                 messages with respect to other processing nodes

17                 over said network;

18                 a local processor;

19                 at least one private write-through cache;

20                 a section of shared memory organized into a

21                 plurality of cache lines, each cache line

22                 including one or more addressable memory

23                 locations;

24                 a cache coherency directory for tracking which of

25                 said nodes have copies of each cache line;

26                 said local processor at a first processing node being

27                 operable for writing data to said private cache at said

28                 first node, as the same data is written to either

29                 shared memory at said first node or sent over said

30                 network for writing to the shared memory and private

31                 cache of a second processing node.

1           3.           The shared memory parallel processing system of

2           claim 2, wherein said section of shared memory is divided

3           into first and second portions, said first portion for

4           storing unchangeable data, and said second portion for

5           storing changeable data.

660760-4954660

1 4. The shared memory parallel processing system of  
2 claim 3, said cache coherency directory for this processing  
3 node listing which nodes of the plurality of nodes have  
4 accessed copies of said cache lines in said second portion  
5 of shared memory at this processing node.

1 5. The shared memory parallel processing system of  
2 claim 4, wherein each said processing node is operable for  
3 reading, storing, and invalidating the shared memory at any  
4 of said plurality of processing nodes selectively by  
5 transmitting and receiving messages over said network, a  
6 first message type for requesting the read of a cache line,  
7 a second message type for returning the requested cache  
8 line, a third message type for storing a cache line, and a  
9 fourth message type for invalidating a cache line.

660760-4954650

1        6.        The shared memory parallel processing system of  
2        claim 5, said network adapter further comprising:

3            a first buffer for transmitting to said network shared  
4            memory read command messages of said first message type  
5            and said second message type;

6            a second buffer for transmitting to said network shared  
7            memory store command messages of said third message  
8            type;

9            a third buffer for transmitting to said network  
10            invalidate messages for said cache coherency directory  
11            of said fourth message type;

12           a fourth buffer for receiving from said network shared  
13           memory read command messages of said first message type  
14           and said second message type;

15           a fifth buffer for receiving from said network shared  
16           memory store command messages of said third message  
17           type; and

18           a sixth buffer for receiving from said network  
19           invalidate messages for said cache coherency directory  
20           of said fourth message type.

1 7. A shared memory parallel processing system,  
2 comprising:

3 a plurality of nodes, each node including a node  
4 memory, at least one cache, and a memory controller;

5 a multi-stage switching network for  
6 interconnecting said processing nodes, said switching  
7 network including a plurality of self-routing switches  
8 cascaded into first, middle and last stages, each said  
9 switch including a plurality of switch inputs and a  
10 plurality of switch outputs, each of said switch  
11 outputs of each said switch coupled to a different  
12 switch input of others of said switches, switch outputs  
13 of said last stage switches including network output  
14 ports, and switch inputs of said first stage switches  
15 comprising network input ports;

16 a system memory distributed to said node memories  
17 of said plurality of nodes and accessible by any node;  
18 each said node memory being organized into a plurality  
19 of addressable word locations;

20 said memory controller at this node operable for  
21 performing local memory access to the portion of system  
22 memory at this node and for performing remote memory  
23 access over said network to the portion of system  
24 memory at other nodes; and

25 a cache coherency controller at this node being  
26 responsive to both local memory accesses and remote  
27 memory accesses to data stored in a word location of  
28 said node memory at this node for caching accessed data  
29 in the cache of this node and for communicating data  
30 for assuring cache coherency throughout said system  
31 over said network.

1 8. The shared memory processing system of claim 7,  
2 said system memory being distributed in equal portions to  
3 each said node memory; and said node memory being further  
4 sub-divided into a first memory section for storing data  
5 that is changeable and a second memory section for storing  
6 data that is unchangeable.

1 9. The shared memory processing system of claim 7,  
2 further comprising node indicia for uniquely identifying  
3 each node.

1 10. The shared memory processing system of claim 7,  
2 said cache coherency controller further comprising:

3 an invalidation directory for storing a list of node  
4 indicia identifying those nodes having accessed a copy  
5 of each said cache line of node memory since the last  
6 time the cache line was changed.

1 11. The shared memory processing system of claim 10,  
2 said cache coherency controller further comprising:

3 an overflow directory for expanding said invalidation  
4 directory when the list of node indicia for a cache  
5 line becomes too long to contain entirely with said  
6 invalidation directory.

650760-1954660

1 12. A shared memory parallel processing system,  
2 comprising:

3 a plurality of nodes, each node including a node  
4 memory, at least one cache, and a memory controller;

5 a multi-stage switching network for interconnecting  
6 said processing nodes, said switching network including  
7 a plurality of self-routing switches cascaded into  
8 first, middle and last stages, each said switch  
9 including a plurality of switch inputs and a plurality  
10 of switch outputs, each of said switch outputs of each  
11 said switch coupled to a different switch input of  
12 others of said switches, switch outputs of said last  
13 stage switches including network output ports, and  
14 switch inputs of said first stage switches comprising  
15 network input ports; and

16 a network adapter responsive to a node connection  
17 request for establishing a connection path to a target  
18 node, first by attempting to establish a quick  
19 connection path across a plurality of segments of said  
20 switching network to said target node, and upon  
21 determining any one of said plurality of segments is  
22 not available, issuing a camp-on connection request to  
23 said target node.

091099-4954660

1 13. The shared memory parallel processing system of  
2 claim 12, further comprising:

3 said plurality of nodes each coupled to one of the  
4 network output ports and to one of the network input  
5 ports;

6 each node further including:

7 receive means for receiving a data message; and

8 send means for sending a data message across an  
9 n-stage switching network from a local node to a  
10 remote node, said send means generating said  
11 connection request including n sequential  
12 connection commands, each sequential connection  
13 command selecting one of said plurality of  
14 connection segments for each of the n switch  
15 stages of said network.

1 14. The shared memory parallel processing system of  
2 claim 12, each said switch being responsive to node  
3 connection requests and camp-on connection requests for  
4 establishing connection segments from any switch input port  
5 to any switch output ports.



093945160750-49546260

1 15. The shared memory parallel processing system of  
2 claim 14, each said switch further comprising:

3 a data bus for transferring said data message;

4 a rejection control line for signalling back to a  
5 sending node a rejection of any connection request;

6 an acceptance control line for signalling back to said  
7 sending node the acceptance of a camp-on connection  
8 request;

9 a valid control line for receiving from said sending  
10 node the activation of a node connection request; and

11 a camp-on control line for receiving from said sending  
12 node the activation of a camp-on connection request.





0345678910111213141516171819

18. The network adapter of claim 17, said send buffers further comprising:

a read send FIFO for storing and forwarding read request messages and response messages from said local node to said remote node;

a store send FIFO for storing and forwarding store messages from said local node to said remote node; and

an invalidation send FIFO for storing and forwarding invalidation messages from said local node to said remote node;

and said receive buffers further comprising:

a read receive FIFO for storing and forwarding read request messages and response messages from said remote node to said local node;

a store receive FIFO for storing and forwarding store messages from said remote node to said local node; and

an invalidation receive FIFO for storing and forwarding invalidation messages from said remote node to said local node.

660760-4954660

1 19. The network adapter of claim 18, further  
2 comprising:

3 a send FIFO selection means for prioritizing the  
4 selection of a data message from one of said three send  
5 FIFO means for transmission to said network by first  
6 selecting data messages from said invalidation send  
7 FIFO and thereafter alternatively selecting data  
8 messages from said read and store send FIFOs;

9 a receive FIFO selection means responsive to said  
10 message type differentiation indicia for selecting one  
11 of said three receive FIFO means for storing a data  
12 message received from said network; and

13 said network adapter being responsive to a node  
14 connection request for establishing a connection path  
15 to a target node, first by attempting to establish a  
16 quick connection path across a plurality of segments of  
17 said switching network to said target node, and upon  
18 determining any one of said plurality of segments is  
19 not available, issuing a camp-on connection request to  
20 said target node.

1 20. A memory controller for a local node of a shared  
2 memory parallel processing system, said node including a  
3 node processor, a node memory, a node cache and an I/O  
4 adapter, said system including a multi-stage switching  
5 network for communications amongst said local node and a  
6 plurality of remote nodes, said node memory including a  
7 changeable portion and an unchangeable portion; said memory  
8 controller comprising:

9 first means responsive to a request by said processor  
10 for access to a memory word for first accessing said  
11 node cache of said local node; and

12 second means responsive to said first means being  
13 unable to access said memory word in said node cache  
14 for accessing said memory word selectively from a cache  
15 line in said node memory or remote memory and storing  
16 said cache line to said node cache.

1 21. The memory controller of claim 20, further  
2 comprising:

3 remote fetch interrupt means for issuing an interrupt  
4 signal to said node processor upon determining that a  
5 requested memory word is located in remote memory for  
6 causing said node processor to switch from a first  
7 instruction stream thread to a second instruction  
8 stream thread.

1 22. The memory controller of claim 20, further  
2 comprising:

3 data message generation means responsive to a request  
4 from a remote node for accessing a cache line  
5 identified by a remote request read address for  
6 generating a read response message to return the  
7 accessed cache line to said remote node, said read  
8 response message including a message header comprising

9 message differentiation indicia for defining said  
10 read request message type;

11 destination node indicia equal to the sector  
12 segment of said node memory for said addressed  
13 memory word;

14 source node indicia set to the node ID number of  
15 the local node;

16 message length indicia for defining said read  
17 request message as being comprised of said message  
18 header only; and

19 memory address indicia for specifying the memory  
20 address of said memory word;

21 said data message generation means further operable for  
22 delivering said read response message to a read send  
23 FIFO of said network adapter for transmission to said  
24 network and the remote node selected by said  
25 destination node indicia.

1 23. The memory controller of claim 20, further  
2 comprising:

3 an invalidation directory;

4 cast-out means for deleting a cache line from said node  
5 cache when said cache is full to provide space for a  
6 new cache line to be stored to said cache; and for  
7 sending the address of the deleted cache line to said  
8 invalidation directory to indicate said node no longer  
9 has a copy of said cache line.

1 24. The memory controller of claim 23, further  
2 comprising:

3 cast-out message generation means responsive to said  
4 cast-out means deleting a cache line addressed to a  
5 remote node for generating a cast-out message to said  
6 remote node to send the cast-out address and the local  
7 node ID number to said remote node over said network;

8 cast-out message receiving means for delivering a  
9 cast-out address and the source node ID number from the  
10 message header of a cast-out message to said  
11 invalidation directory.







1 29. A method for operating bi-directional network  
2 adapter for interfacing a local node of a shared memory  
3 parallel processing system to a multi-stage switching  
4 network for communications with a remote node, each said  
5 node including a node memory including a changeable portion  
6 and an unchangeable portion, and a node cache; comprising  
7 the steps of:

8 operating a plurality of send buffers for storing and  
9 forwarding data messages from said local node to said  
10 remote node over said network, and

11 operating a plurality of receive buffers for storing  
12 and forwarding a plurality of data messages from said  
13 remote node to said local node over said multi-stage  
14 network;

15 said data messages including:

16 an invalidation message for invalidating a cache  
17 line that was accessed by a remote node after said  
18 cache line has changed;

19 a read request message for requesting access of a  
20 cache line from a remote node;

21 a response message for returning a cache line over  
22 the network to a remote node that has previously  
23 requested data by a read request message; and

24 a store message storing a changed cache line to a  
25 remote node.

093454-091099  
660760-1554650

1 30. The method of claim 29, further comprising the  
2 steps of:

3 operating a read send FIFO for storing and forwarding  
4 read request messages and response messages from said  
5 local node to said remote node;

6 operating a store send FIFO for storing and forwarding  
7 store messages from said local node to said remote  
8 node; and

9 operating an invalidation send FIFO for storing and  
10 forwarding invalidation messages from said local node  
11 to said remote node;

12 operating a read receive FIFO for storing and  
13 forwarding read request messages and response messages  
14 from said remote node to said local node;

15 operating a store receive FIFO for storing and  
16 forwarding store messages from said remote node to said  
17 local node; and

18 operating an invalidation receive FIFO for storing and  
19 forwarding invalidation messages from said remote node  
20 to said local node.

add 15